(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)特許出願公開番号
特開2005-182696
(P2005-182696A)
(43)公開日　平成17年7月7日(2005.7.7)

(51)Int.Cl.[7]	F I		テーマコード（参考）
G06N 3/00	G06N 3/00	550Z	5B075
G06F 17/30	G06F 17/30	210D	

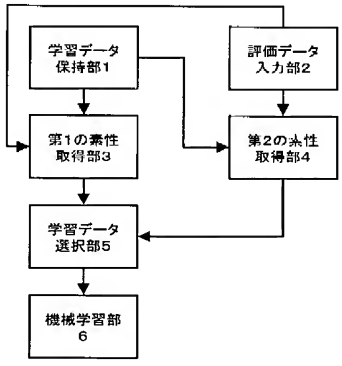
審査請求 未請求　請求項の数 21　OL　（全 17 頁）

(21)出願番号　特願2003-426330(P2003-426330)
(22)出願日　平成15年12月24日(2003.12.24)

(71)出願人　000005496
富士ゼロックス株式会社
東京都港区赤坂二丁目１７番２２号
(74)代理人　100086531
弁理士　澤田　俊夫
(74)代理人　100093241
弁理士　宮田　正昭
(74)代理人　100101801
弁理士　山田　英治
(72)発明者　吉村　宏樹
神奈川県足柄上郡中井町境４３０　グリーンテクなかい　富士ゼロックス株式会社内
(72)発明者　増市　博
神奈川県足柄上郡中井町境４３０　グリーンテクなかい　富士ゼロックス株式会社内
最終頁に続く

(54)【発明の名称】機械学習システム及び機械学習方法、並びにコンピュータ・プログラム

(57)【要約】
【課題】　評価データに適切な学習データを用いて精度の高い機械学習を行なう。
【解決手段】　評価対象となるデータが与えられる度に、学習データと評価データから学習データを選択するための第２の素性を抽出し、第２の素性を基に機械学習に用いるのに適した学習データを選択することができる。すなわち、評価データに適した学習データを用いて機械学習を行なうことが可能となる。また、評価データ毎に学習データを取捨選択するので、素性数の計算上の限界を超える素性を用意しておくことができる。
【選択図】　図１



【特許請求の範囲】
【請求項１】
データの評価があらかじめ付与されている教師ありの機械学習を行なう機械学習システムであって、
機械学習を行なうための学習データの候補を評価とともに保持する学習データ保持部と、
評価対象となるデータを受け取る評価データ入力部と、
前記学習データ保持部に保持されるデータ及び前記評価データ入力部で受け取ったデータから、機械学習及び評価を行なう際に用いる第１の素性情報を抽出する第１の素性取得部と、
前記学習データ保持部に保持されるデータ及び前記評価データ入力部で受け取ったデータから、機械学習を行なう際に用いる学習データを選択するための第２の素性情報を抽出する第２の素性取得部と、
前記第２の素性取得部から得られる第２の素性情報に基づいて、前記学習データ保持部に保持されている学習データの候補の中から機械学習を行なう際に用いる学習データを選択する学習データ選択部と、
前記学習データ選択部によって選択された各学習データの評価と、前記第１の素性取得部から得られた各データの第１の素性情報を基に、素性とその評価の間の対応関係を学習する機械学習部と、
を具備することを特徴とする機械学習システム。
【請求項２】
前記学習データ保持部は、自然言語文からなるテキスト・データを保持し、
前記第１の素性取得部及び前記第２の素性取得部は、形態素解析処理又は構文解析処理により学習データから素性情報を取得する、
ことを特徴とする請求項１に記載の機械学習システム。
【請求項３】
前記機械学習部は、ベクトル空間法に基づいてテキスト・データの素性と評価の間の対応規則を計算する、
ことを特徴とする請求項２に記載の機械学習システム。
【請求項４】
前記機械学習部は、Ｓｕｐｐｏｒｔ　Ｖｅｃｔｏｒ　Ｍａｃｈｉｎｅに基づいてテキスト・データの素性と評価の間の対応規則を計算する、
ことを特徴とする請求項２に記載の機械学習システム。
【請求項５】
前記学習データ選択部は、前記機械学習部が機械学習を行なう際に用いる素性の数に基づいて学習データを選択する、
ことを特徴とする請求項１に記載の機械学習システム。
【請求項６】
前記学習データ選択部は、前記機械学習部が機械学習を行なう際に用いる学習データの数に基づいて学習データを選択する、
ことを特徴とする請求項１に記載の機械学習システム。
【請求項７】
前記学習データ選択部は、前記機械学習部が機械学習を行なう際に用いる素性の種類に基づいて学習データを選択する、
ことを特徴とする請求項１に記載の機械学習システム。
【請求項８】
前記学習データ選択部は、前記機械学習部が機械学習を行なう際に用いる学習データの種類に基づいて学習データを選択する、
ことを特徴とする請求項１に記載の機械学習システム。
【請求項９】

前記機械学習部は、前記第２の素性取得部から得られる各データの素性に基づいて機械学習を行ない、
前記学習データ選択部は、機械学習結果を学習データの選択に用いる、
ことを特徴とする請求項１に記載の機械学習システム。
【請求項１０】
前記学習データ選択部は、Ｎ回以上の素性の取得に基づいて学習データを選択する、
ことを特徴とする請求項１に記載の機械学習システム。
【請求項１１】
機械学習を行なうためにあらかじめ評価とともに保持されている学習データの候補を用いて教師ありの機械学習を行なう機械学習方法であって、
すべての学習データについて第１の素性情報及び第２の素性情報を取得するステップと、
学習データについての評価と第１の素性情報及び第２の素性情報との関係を学習するステップと、
評価データについての第２の素性情報を取得するステップと、
第２の素性情報を基に評価データの評価を行なうステップと、
評価のよい第２の素性情報を基に学習データを選択するステップと、
選択された学習データを用いて第１の素性情報を基に評価データの評価を行なうステップと、
を具備することを特徴とする機械学習方法。
【請求項１２】
前記学習データとして自然言語文からなるテキスト・データを用い、
前記第１の素性取得ステップ及び前記第２の素性取得ステップでは、形態素解析処理又は構文解析処理により学習データから素性情報を取得する、
ことを特徴とする請求項１１に記載の機械学習方法。
【請求項１３】
前記機械学習ステップでは、ベクトル空間法に基づいてテキスト・データの素性と評価の間の対応規則を計算する、
ことを特徴とする請求項１２に記載の機械学習方法。
【請求項１４】
前記機械学習ステップでは、Ｓｕｐｐｏｒｔ Ｖｅｃｔｏｒ Ｍａｃｈｉｎｅに基づいてテキスト・データの素性と評価の間の対応規則を計算する、
ことを特徴とする請求項１２に記載の機械学習方法。
【請求項１５】
前記学習データ選択ステップでは、前記機械学習ステップにおいて機械学習を行なう際に用いる素性の数に基づいて学習データを選択する、
ことを特徴とする請求項１１に記載の機械学習方法。
【請求項１６】
前記学習データ選択ステップでは、前記機械学習ステップにおいて機械学習を行なう際に用いる学習データの数に基づいて学習データを選択する、
ことを特徴とする請求項１１に記載の機械学習方法。
【請求項１７】
前記学習データ選択ステップでは、前記機械学習ステップにおいて機械学習を行なう際に用いる素性の種類に基づいて学習データを選択する、
ことを特徴とする請求項１１に記載の機械学習方法。
【請求項１８】
前記学習データ選択ステップでは、前記機械学習ステップにおいて機械学習を行なう際に用いる学習データの種類に基づいて学習データを選択する、
ことを特徴とする請求項１１に記載の機械学習方法。
【請求項１９】

前記機械学習ステップでは、前記第２の素性取得ステップにおいて得られる各データの素性に基づいて機械学習を行ない、
前記学習データ選択ステップでは、機械学習結果を学習データの選択に用いる、
ことを特徴とする請求項１１に記載の機械学習方法。
【請求項２０】
前記学習データ選択ステップでは、Ｎ回以上の素性の取得に基づいて学習データを選択する、
ことを特徴とする請求項１１に記載の機械学習方法。
【請求項２１】
機械学習を行なうためにあらかじめ評価とともに保持されている学習データの候補を用いて教師ありの機械学習を行なうための処理をコンピュータ・システム上で実行するようにコンピュータ可読形式で記述されたコンピュータ・プログラムであって、
すべての学習データについて第１の素性情報及び第２の素性情報を取得するステップと、
学習データについての評価と第１の素性情報及び第２の素性情報との関係を学習するステップと、
評価データについての第２の素性情報を取得するステップと、
第２の素性情報を基に評価データの評価を行なうステップと、
評価のよい第２の素性情報を基に学習データを選択するステップと、
選択された学習データを用いて第１の素性情報を基に評価データの評価を行なうステップと、
を具備することを特徴とするコンピュータ・プログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、学習データを入力とし、統計処理手法を用いることによって、データの特徴を説明するための規則を出力する機械学習システム及び機械学習方法、並びにコンピュータ・プログラムに係り、特に、学習データ中の各データに、データの評価があらかじめ（人手によって）付与されている教師ありの機械学習を行なう機械学習システム及び機械学習方法、並びにコンピュータ・プログラムに関する。
【０００２】
さらに詳しくは、本発明は、教師あり機械学習により、学習データ中の素性とその評価の間の対応規則を学習することによって、非学習データが与えられたときにその評価を予測する機械学習システム及び機械学習方法、並びにコンピュータ・プログラムに係り、特に、非学習データに適切な学習データを用いて精度の高い機械学習を行なう機械学習システム及び機械学習方法、並びにコンピュータ・プログラムに関する。
【背景技術】
【０００３】
昨今の情報処理技術の発展と普及とも相俟って、産業活動や日常生活におけるさまざまな処理や作業の自動化が進められている。ここで、機械を自動化するには、さまざまなパラメータを決定する必要がある。このようなパラメータを機械自身で決定するために、いわゆる「機械学習」が導入されている。
【０００４】
機械学習では、学習データを入力とし、統計処理手法を用いることによって、データの特徴を説明するための規則を出力する。例えば、機械自身がある動作を行なったときに得られた結果を学習データとして入力してこれを統計的に評価し、その評価を自分自身の行動決定パラメータに反映させる。
【０００５】
機械自身が自分で評価できない場合には、「教師あり学習」と呼ばれる方法により、人間が期待する解を機械に与え、その解に至るように学習のパラメータを調整する。これに

対し、自分で評価することを「教師なし学習」と呼ぶ。教師あり学習として、ニューラル・ネットワークなどを利用した処理方法が挙げられる。また、教師なし学習として、ＥＭ（ｅｘｐｅｃｔａｔｉｏｎ　ｍａｘｉｍｉｚａｔｉｏｎ）アルゴリズムなどを利用した処理方法が挙げられる。

【０００６】

前者の教師あり機械学習では、学習データ中の各データに、データの評価があらかじめ（人手によって）付与されている。学習データ中の各データの特徴（以下、「素性」とも呼ぶ）とその評価の間の対応関係（対応規則）を学習することによって、非学習データ（テスト・データ）が与えられたときにその評価を予測することが可能となる。

【０００７】

現在、Ｓｕｐｐｏｒｔ　Ｖｅｃｔｏｒ　Ｍａｃｈｉｎｅ（ＳＶＭ）や、Ｍａｘｉｍｕｍ　Ｅｎｔｏｒｐｙ（ＭＥ）などさまざまな教師あり機械学習手法が提案され、自然言語処理や生物情報学などのさまざまな分野で実用的に使用されている。機械学習手法の詳細については、例えば、Ｆａｂｒｉｚｉｏ　Ｓｅｂａｓｔｉａｎｉ著“Ｍａｃｈｉｎｅ　Ｌｅａｒｎｉｎｇ　ｉｎ　Ａｕｔｏｍａｔｅｄ　Ｔｅｘｔ　Ｃａｔｅｇｏｒｉｚａｔｉｏｎ”（ＡＣＭ　Ｃｏｍｐｕｒｔｉｎｇ　Ｓｕｒｖｅｙｓ　Ｖｏｌ．３４，　Ｎｏ．１，ｐｐ．１－４７，２００２）に記載されている。

【０００８】

例えば、異常（不正）であるか否かがわかっていないデータ（教師なしデータ）を基に、異常データを特徴付けるルールを生成し、さらに得られたルールを用いて効率よく異常なデータを検出することができる（例えば、特許文献１を参照のこと）。すなわち、データ集合内にある異常なデータを特徴付けるルールを生成する外れ値検出するために、異常であることの度合いを示す外れ値度を算出し、かつ外れ値度に基づいてサンプリングすることにより、異常なデータであるか否かを示すラベルを付与した各データの集合に基づく教師あり学習により、異常なデータを特徴付けるルールを生成する教師あり学習部を備え、効率よく異常なデータを検出することができる。

【０００９】

【特許文献１】特開２００３－５９７０号公報

【発明の開示】

【発明が解決しようとする課題】

【００１０】

従来の教師あり機械学習では、なるべく多くの種類の素性をあらかじめ設定しておき、それら大量の素性を用いて機械学習を行なう（すなわち、対応規則を生成する）。しかしながら、実際に扱うことが可能な素性の（種類の）数には限界があり、限界を超えた場合には計算量が増大し、学習を実行することが不可能となる。また、素性と評価の間の対応規則と無関係な素性を大量に学習に含めると、得られる対応規則の信頼性を損ねることになる。

【００１１】

このように、適切な素性の選択は、精度の高い機械学習を実現するために不可欠の要素である。しかしながら、適切な素性であるか否かは、どのようなテスト・データが入力されるかに依存するため、あらゆるテスト・データに平均的に有効な素性を、素性数の計算上の上限までの範囲で用意しておくしかなかった。

【００１２】

例えば前述した特許文献１においては、大多数のデータが従う確率分布から外れたデータ（発生しにくいデータ）を、「統計的外れ値」としてそのデータを異常（不正）と同定することができるが、評価データに対し相対的に適した学習データを用いるものではなく、上述したような技術的課題を解決していない。

【００１３】

本発明の目的は、学習データ中の各データに、データの評価があらかじめ（人手によって）付与されている教師ありの機械学習を高い精度で行なうことができる、優れた機械学

習システム及び機械学習方法、並びにコンピュータ・プログラムを提供することにある。

【００１４】

本発明のさらなる目的は、非学習データに適切な学習データを用いて精度の高い機械学習を行なうことができる、優れた機械学習システム及び機械学習方法、並びにコンピュータ・プログラムを提供することにある。

【課題を解決するための手段】

【００１５】

本発明は、上記課題を参酌してなされたものであり、その第１の側面は、データの評価があらかじめ付与されている教師ありの機械学習を行なう機械学習システムであって、

機械学習を行なうための学習データの候補を評価とともに保持する学習データ保持部と、

評価対象となるデータを受け取る評価データ入力部と、

前記学習データ保持部に保持されるデータ及び前記評価データ入力部で受け取ったデータから、機械学習及び評価を行なう際に用いる第１の素性情報を抽出する第１の素性取得部と、

前記学習データ保持部に保持されるデータ及び前記評価データ入力部で受け取ったデータから、機械学習を行なう際に用いる学習データを選択するための第２の素性情報を抽出する第２の素性取得部と、

前記第２の素性取得部から得られる第２の素性情報に基づいて、前記学習データ保持部に保持されている学習データの候補の中から機械学習を行なう際に用いる学習データを選択する学習データ選択部と、

前記学習データ選択部によって選択された各学習データの評価と、前記第１の素性取得部から得られた各データの第１の素性情報を基に、素性とその評価の間の対応関係を学習する機械学習部と、

を具備することを特徴とする機械学習システムである。

【００１６】

図１には、本発明に係る機械学習システムの機能構成を模式的に示している。同図に示すように、機械学習システムは、学習データ保持部１と、評価データ入力部２と、第１の素性取得部３と、第２の素性取得部と、学習データ選択部５と、機械学習部６で構成される。

【００１７】

第１の素性取得部３は、学習データ保持部１に保持されるデータ及び評価データ入力部２で受け取ったデータから、機械学習並びに評価を行なう際に用いる第１の素性情報を抽出する。

【００１８】

これに対し、第２の素性取得部４は、学習データ保持部１に保持されるデータ及び評価データ入力部２で受け取ったデータから、機械学習を行なう際に用いる学習データを選択するための第２の素性情報を抽出する。さらに、学習データ選択部５は、第２の素性取得部４から得られる素性に基づいて、機械学習を行なう際に用いる学習データを選択する。

【００１９】

そして、機械学習部６は、学習データ選択部５によって選択された各学習データの評価と、第１の素性取得部３から得られた各データの素性を基に、素性とその評価の間の対応関係を計算する。

【００２０】

本発明に係る機械学習システムでは、学習時には、すべての学習データについて第１及び第２の素性情報を取得し、第１及び第２の素性情報を基にそれぞれ学習する。そして、評価時には、まず評価データについての第２の素性情報を取得し、第２の素性情報を基に評価データの評価を行ない、次いで、評価のよい第２の素性情報を基に学習データを選択し、選択された学習データを用いて第１の素性情報を基に評価データの評価を行なう。

【００２１】

本発明に係る機械学習システムは、素性情報を分離することにより、機械学習の高精度化を図るものである。すなわち、評価対象となるデータが与えられる度に、学習データと評価データから学習データを選択するための第２の素性を抽出し、第２の素性を基に機械学習に用いるのに適した学習データを選択することができる。すなわち、評価データに適した学習データを用いて機械学習を行なうことが可能となる。また、評価データ毎に学習データを取捨選択するので、素性数の計算上の限界を超える素性を用意しておくことができる。

【００２２】

ここで、前記学習データ保持部は、例えば、自然言語文からなるテキスト・データを保持する。そして、前記第１の素性取得部及び前記第２の素性取得部は、形態素解析処理又は構文解析処理により、学習データから形態素や構文解析儀などを素性情報として取得することができる。

【００２３】

本発明に係る機械学習システムは、例えば文書分類システムに適用される。そして、前記機械学習部は、例えば、ベクトル空間法に基づいてテキスト・データの素性と評価の間の対応規則を計算することができる。

【００２４】

ここで言うベクトル空間法とは、全テキスト・データに含まれる全単語のうち出現頻度の多い所定数のものを「特徴表現語」として抽出し、各単語と特徴表現語が共起（同じテキスト・データで出現）する回数を共起行列として表した単語ベクトルを生成し、次いで、対象とするテキスト・データに含まれる全単語の単語ベクトルの総和を正規化した文書ベクトルを生成し、評価対象となるテキスト・データについても同様の評価文書ベクトルを生成し、文書ベクトルに基づいて評価を行なう方法である。各分類の文書ベクトルと評価文書ベクトルとの内積により、評価対象のテキスト・データを分類することができる。

【００２５】

あるいは、前記機械学習部は、Ｓｕｐｐｏｒｔ Ｖｅｃｔｏｒ Ｍａｃｈｉｎｅに基づいてテキスト・データの素性と評価の間の対応規則を計算するようにしてもよい。Ｓｕｐｐｏｒｔ Ｖｅｃｔｏｒ Ｍａｃｈｉｎｅは、ノンパラメトリックなパターン分類器の１つであり、学習の最適解として求められた分離超平面による線形識別を行ない、学習資料を線形分離することが不適切な場合には学習資料を元のパターン空間からより高次のパターン空間に非線形写像し高次元空間で分離超平面を構築し線形識別を行なうことができる。

【００２６】

また、前記学習データ選択部は、前記機械学習部が機械学習を行なう際に用いる素性の数に基づいて学習データを選択するようにしてもよい。

【００２７】

あるいは、前記学習データ選択部は、前記機械学習部が機械学習を行なう際に用いる学習データの数に基づいて学習データを選択するようにしてもよい。

【００２８】

あるいは、前記学習データ選択部は、前記機械学習部が機械学習を行なう際に用いる素性の種類に基づいて学習データを選択するようにしてもよい。

【００２９】

あるいは、前記学習データ選択部は、前記機械学習部が機械学習を行なう際に用いる学習データの種類に基づいて学習データを選択するようにしてもよい。

【００３０】

また、前記機械学習部は、前記第２の素性取得部から得られる各データの素性に基づいて機械学習を行ない、前記学習データ選択部は、機械学習結果を学習データの選択に用いるようにしてもよい。

【００３１】

また、前記学習データ選択部は、Ｎ回以上の素性の取得に基づいて学習データを選択するようにしてもよい。

【００３２】
また、本発明の第２の側面は、機械学習を行なうためにあらかじめ評価とともに保持されている学習データの候補を用いて教師ありの機械学習を行なうための処理をコンピュータ・システム上で実行するようにコンピュータ可読形式で記述されたコンピュータ・プログラムであって、
すべての学習データについて第１の素性情報及び第２の素性情報を取得するステップと、
学習データについての評価と第１の素性情報及び第２の素性情報との関係を学習するステップと、
評価データについての第２の素性情報を取得するステップと、
第２の素性情報を基に評価データの評価を行なうステップと、
評価のよい第２の素性情報を基に学習データを選択するステップと、
選択された学習データを用いて第１の素性情報を基に評価データの評価を行なうステップと、
を具備することを特徴とするコンピュータ・プログラムである。
【００３３】
本発明の第２の側面に係るコンピュータ・プログラムは、コンピュータ・システム上で所定の処理を実現するようにコンピュータ可読形式で記述されたコンピュータ・プログラムを定義したものである。換言すれば、本発明の第２の側面に係るコンピュータ・プログラムをコンピュータ・システムにインストールすることによって、コンピュータ・システム上では協働的作用が発揮され、本発明の第１の側面に係る機械学習システムと同様の作用効果を得ることができる。
【発明の効果】
【００３４】
本発明によれば、学習データ中の各データに、データの評価があらかじめ（人手によって）付与されている教師ありの機械学習を高い精度で行なうことができる、優れた機械学習システム及び機械学習方法、並びにコンピュータ・プログラムを提供することができる。
【００３５】
また、本発明によれば、評価データに適切な学習データを用いて精度の高い機械学習を行なうことができる、優れた機械学習システム及び機械学習方法、並びにコンピュータ・プログラムを提供することができる。
【００３６】
本発明に係る機械学習システムによれば、テスト・データが与えられる度に適切な学習データを選択することによって、テスト・データに適した機械学習を行なうことが可能となり、また、素性数の計算上の上限を超える素性を用意しておくことができる。
【００３７】
本発明のさらに他の目的、特徴や利点は、後述する本発明の実施形態や添付する図面に基づくより詳細な説明によって明らかになるであろう。
【発明を実施するための最良の形態】
【００３８】
以下、図面を参照しながら本発明の実施形態について詳解する。
【００３９】
図２には、本発明の一実施形態に係る機械学習システムの機能構成を模式的に示している。図示の機械学習システムは、学習コーパス保持部１１と、評価データ入力部１２と、形態素解析部１３と、文書長さ取得部１４と、学習コーパス選択部１５と、単語ベクトル生成部１６と、文書ベクトル生成部１７と、評価データ分類部１８で構成され、機械学習手段としてベクトル空間法を採用する。この機械学習システムは、実際には、パーソナル・コンピュータのような一般的な計算機システムに所定の機械学習アプリケーションを実行するという形態で実現される。

【００４０】
なお、以下に説明する本発明の実施形態は、機械学習手法を新聞記事の分類（「政治経済」分野の記事であるか「スポーツ」分野の記事であるか）などの文書分類システムに応用する場合を例に挙げているが、統計処理に基づく教師あり機械学習手法を用いるものであれば、アンケート分類及び質問応答など分類を要するあらゆる分野への応用であっても同様の効果を得ることが可能である。その他、テキスト分類のみならず数値データを含む分類や画像の分類など、いかなる機械学習手法を用いるものであっても、同様の効果を得ることが可能である。
【００４１】
学習コーパス保持部１１は、学習コーパスとしての複数の新聞記事を、記事毎に「政治経済」分野に属するか「スポーツ」分野に属するかを人手で判断した評価結果とともに、コンピュータ内部に保持する。
【００４２】
評価データ入力部１２は、単一の入力された新聞記事テキストが、「政治経済」の分野に属する記事であるか「スポーツ」の分野に属する記事であるかを判断するために、単一の新聞記事テキストを評価データとして受け取る。
【００４３】
形態素解析部１３は、学習コーパス保持部１１に保持されているすべての新聞記事テキスト、及び評価データ入力部１２に入力された単一の新聞記事テキストに対してそれぞれ形態素解析処理を施し、これらの新聞記事テキストを単語へと分割し、第１の素性情報としての形態素解析結果を取得する。
【００４４】
文書長さ取得部１４は、学習コーパス保持部１１に保持されているすべての新聞記事テキスト及び評価データ入力部１２に入力された新聞記事テキストに対して、各テキストの長さ（テキスト中に含まれる文字数）を計測し、これを第２の素性情報として取得する。
【００４５】
学習コーパス選択部１５は、文書長さ取得部１４から得られるテキストの長さを基に、学習コーパスから、実際に機械学習で用いる学習データを選択する。すなわち、学習コーパス選択部１５は、まず評価データについての第２の素性情報としての文書長さを取得し、第２の素性情報を基に評価データの評価を行ない、次いで、評価のよい第２の素性情報を基に学習データを選択する。ここで、Ｌは評価データ入力部１２に入力された新聞記事テキストの長さとし、Ｔはあらかじめ設定された正の定数とした場合、Ｌ－ＴよりもＬ、Ｌ＋Ｔよりも短いテキストを、学習コーパス保持部１１に保持されている新聞記事テキストから選択する。
【００４６】
学習コーパス選択部１５によって選択された新聞記事テキストを用いて、機械学習、並びに評価データ入力部１２から入力された新聞記事テキストの評価を行なう。本実施形態では、機械学習を文書分類システムに応用するが、ベクトル空間法に基づいて、テキスト・データの素性と評価の間の対応規則を計算する。図示の例では、機械学習手段は、単語ベクトル生成部１６と、文書ベクトル生成部１７と、評価データ分類部１８で構成される。
【００４７】
単語ベクトル生成部１６は、テキスト中の各単語に対して、対応する多次元ベクトル（単語ベクトル）を計算する。以下、単語ベクトルを計算するアルゴリズムについて説明する。
【００４８】
ステップ１：
学習コーパス選択部１５によって選択された新聞記事テキストを対象として、形態素解析部１２で得られた全単語のうち、出現頻度の多いものから順にｎ個の単語を選択する。ここで得られたｎ個の単語のことを、以下では「特徴表現語」と呼ぶことにする。ｎの値

は、学習コーパス選択部１５によって選択された全新聞記事テキストに含まれる総異なり単語数の２０％とする。

【００４９】

但し、通常、新聞記事のキーワードとなりにくく且つ文に含まれる単語数の多い「は」又は「が」などの助詞については、ストップ・ワードとして、特徴表現語としてカウントしない場合もある。

【００５０】

ステップ２：

学習コーパス選択部１５によって選択された新聞記事テキストを対象とし、形態素解析部１２から得られた全単語を行とし、ステップ１で得られた特徴表現語を列として構成される行列を作成する。例えば、学習コーパス選択部１５によって選択された新聞記事テキストを対象として形態素解析部１２から得られた全単語の総異なり語数が１０万であれば、ｎの値は２万となり、１０万行×２万列の行列ができることになる。

【００５１】

この行列の各要素には、その要素の行に対応する単語と列に対応する特徴表現語が、新聞記事中で何度共起しているか（すなわち、同じ新聞記事中に同時に何度出現しているか）を記録する。こうして得られた行列のことを「共起行列」と呼ぶことにする。このようにして、学習コーパス選択部１５によって選択された新聞記事中の全単語をそれぞれｎ次元（２万次元）のベクトルで表現する共起行列を作成することができる。このベクトルは、各単語が学習コーパス選択部１５によって選択された新聞記事中で、どのようなコンテキストで出現し易い傾向にあるかを示すベクトルであると言える。

【００５２】

ステップ３：

ステップ２で得られたｎ次元のベクトルは次元数が大きいため、後に必要となる処理で計算時間が膨大なものになってしまう。そこで、計算処理を実時間の範囲に抑えるために、元のｎ次元のベクトルを行列の次元圧縮手法によって、ｎ’次元（数百次元）のベクトルへと圧縮する（ｎ’＜ｎ）。次元圧縮手法にはさまざまなものが存在するが、例えばＢｅｒｒｙ，Ｍ．、Ｄｏ，Ｔ．、Ｏ’Ｂｒｉｅｎ，Ｇ．、Ｋｒｉｓｈｎａ，Ｖ．及びＶａｒａｄｈａｎ，Ｓ．共著“ＳＶＤＰＡＣＫＣ ＵＳＥＲ’Ｓ ＧＵＩＤＥ”（Ｔｅｃｈ．Ｒｅｐ．ＣＳ－９３－１９４．Ｕｎｉｖｅｒｓｉｔｙ ｏｆ Ｔｅｎｎｅｓｓｅｅ，Ｋｎｏｘｖｉｌｌｅ，ＴＮ（１９９３））で詳細な説明がなされている、Ｓｉｎｇｕｌａｒ Ｖａｌｕｅ Ｄｅｃｏｍｐｏｓｉｔｉｏｎ（特異値分解）を利用する手法がその代表例である。このようにして新聞記事中のすべての単語に対して得られたｎ’次元のベクトルのことを「単語ベクトル」と呼ぶことにする。

【００５３】

文書ベクトル生成部１７は、単語ベクトル生成部１６で得られた単語ベクトルを用いて、学習コーパス選択部１５によって選択された各新聞記事についての文書ベクトルを計算する。ここで言う文書ベクトルとは、対象とする新聞記事に含まれる全単語に対応する単語ベクトルの総和を正規化した（ベクトルの長さを１とした）ベクトルのことである。このようにして得られた文書ベクトルは、学習コーパス選択部１５によって選択された新聞記事集合を学習データとし、新聞記事に含まれる特徴表現語を各記事の素性とした場合に得られる機械学習の結果であると言える。

【００５４】

また、同様に、評価データ入力部１２に入力された新聞記事テキストに含まれる全単語に対応する単語ベクトルの総和を正規化した（ベクトルの長さを1とした）ベクトルを生成する（但し、対応する単語ベクトルが存在しない単語は無視する）。この文書ベクトルのことを「評価文書ベクトル」と呼ぶことにする。

【００５５】

評価データ分類部１８は、文書ベクトル生成部１７から得られる各文書ベクトルを参照し、評価データ入力部１２に入力された新聞記事テキストが例えば「政治経済」分野に属

する記事であるか「スポーツ」分野に属する記事であるかを判断する。

【００５６】

まず、文書ベクトル生成部１７から得られる文書ベクトルのうち「政治経済」分野に属する記事に対応する文書ベクトルの総和を計算し正規化する。同様に、「スポーツ」分野に属する記事に対応する文書ベクトルの総和を計算し正規化する。それぞれを「政治経済文書ベクトル」と「スポーツ文書ベクトル」と呼ぶことにする。

【００５７】

次に、「評価文書ベクトル」と「政治経済文書ベクトル」との間の類似度をたとえばベクトルの内積の値として計算し、同様に「評価文書ベクトル」と「スポーツ文書ベクトル」との間の内積を計算する。評価データ入力部１２に入力された新聞記事は、内積の値が大きい方の分野と内容的に近いと判断することが可能であり、与えられた新聞記事を「政治経済」か「スポーツ」のいずれかに分類することが可能である。

【００５８】

このように、学習コーパスから実際に機械学習を行なう新聞記事テキストを選択するための第２の素性（上述した実施形態では「テキストの長さ」）と、機械学習を行なう際に利用する第１の素性（上述した実施形態では「テキスト中の単語の出現頻度」）の両素性を用いることによって、評価データ入力部１２に入力された新聞記事毎に、適切な学習データを利用した機械学習を行なうことが可能である。すなわち、評価データ入力部１２に入力された新聞記事と同程度の長さの新聞記事のみを学習データとすることによって、入力に適した分類が行なうことができる。

【００５９】

新聞記事の長さ毎に単語の出現頻度分布の傾向が異なる場合、評価データ入力部１２に入力された新聞記事の長さを勘案せず学習データ全体を使用すると、評価データ入力部１２に入力された新聞記事の分類を判断する目的に対して不適切な学習データを用いてしまうことになる。

【００６０】

さらに、学習コーパスの一部を用いて機械学習を行なうため、使用する素性の種類（単語の種類）も、学習コーパス全体を用いて機械学習を行なう場合と比較して、軽減することができる。上述した実施形態では、学習コーパス選択部１５が、あらかじめ設定された閾値Ｔを用いて機械学習に利用する新聞記事を選択している。

【００６１】

これに対し、特徴表現語の数ｎ（学習コーパス選択部１５によって選択された全新聞記事テキストに含まれる総異なり単語数の２０％）を、あらかじめ定数として設定しておき、Ｔを変数として、ｎがあらかじめ設定された値となるようにＴを調節することも可能である。また、学習コーパス選択部１５によって選択される記事を、あらかじめ設定された記事数となるように、Ｔを調節することも可能である。

【００６２】

また、上述した実施形態では、機械学習に用いる素性として形態素解析結果から得られる単語の出現頻度を用いたが、本発明の要旨はこれに限定されるものではない。すなわち、機械学習においては、形態素解析結果以外であっても、テキストの特徴を表現し得るものであれば、いかなるものであっても素性となり得る。例えば、形態素解析の代わりに構文解析を施し、新聞記事中において係り受け関係を有する単語のペアの出現頻度を、機械学習及び評価を行なう際に用いる第１の素性とすることも可能である。

【００６３】

また、上述した実施形態では、ベクトル空間法に基づく機械学習手法を用いたが、これをＳｕｐｐｏｒｔ Ｖｅｃｔｏｒ Ｍａｃｈｉｎｅのような他の手法で置き換えることも可能である。ここで、Ｓｕｐｐｏｒｔ Ｖｅｃｔｏｒ Ｍａｃｈｉｎｅは、ノンパラメトリックなパターン分類器の１つであり、学習の最適解として求められた分離超平面による線形識別を行ない、学習資料を線形分離することが不適切な場合には学習資料を元のパターン空間からより高次のパターン空間に非線形写像し高次元空間で分離超平面を構築し線形識

別を行なう。ＳＶＭは、テキスト分類などの分類予測精度が高いとされている機械学習手法であるため、本実施形態の機械学習手段に用いることが可能である。Ｓｕｐｐｏｒｔ Ｖｅｃｔｏｒ Ｍａｃｈｉｎｅの学習結果に基づく分類処理の詳細については、例えば、Ｆａｂｒｉｚｉｏ Ｓｅｂａｓｔｉａｎｉ著“Ｍａｃｈｉｎｅ Ｌｅａｒｎｉｎｇ ｉｎ Ａｕｔｏｍａｔｅｄ Ｔｅｘｔ Ｃａｔｅｇｏｒｉｚａｔｉｏｎ”（ＡＣＭ Ｃｏｍｐｕｒｔｉｎｇ Ｓｕｒｖｅｙｓ Ｖｏｌ．３４，Ｎｏ．１，ｐｐ．１－４７，２００２）などに記載されている。

【００６４】

Ｓｕｐｐｏｒｔ Ｖｅｃｔｏｒ Ｍａｃｈｉｎｅを用いた機械学習では、素性情報は、図３で示すようなデータ集合となる。同図に示す例では、単語Ｗ１という素性が入力されており、文Ｓ１内の単語Ｗ１の個数（１個）がカウントされていることを示している。

【００６５】

図４には、Ｓｕｐｐｏｒｔ Ｖｅｃｔｏｒ Ｍａｃｈｉｎｅを機械学習に適用した場合の機械学習システムの機能構成を模式的に示している。図示の機械学習システムは、学習コーパス保持部１１と、評価データ入力部１２と、形態素解析部１３と、文書長さ取得部１４と、学習コーパス選択部１５と、単語素性生成部２６と、文書素性生成部２７と、評価データ分類部２８で構成される。この機械学習システムは、実際には、パーソナル・コンピュータのような一般的な計算機システムに所定の機械学習アプリケーションを実行するという形態で実現される。以下では、機械学習手法を新聞記事の分類（「政治経済」分野の記事であるか「スポーツ」分野の記事であるか）などの文書分類システムに応用する場合を例に説明する。

【００６６】

学習コーパス保持部１１は、学習コーパスとしての複数の新聞記事を、記事毎に分野を人手で判断した評価結果とともに、コンピュータ内部に保持する。

【００６７】

評価データ入力部１２は、単一の入力された新聞記事テキストを評価データとして受け取る。

【００６８】

形態素解析部１３は、学習コーパス保持部１１に保持されているすべての新聞記事テキスト、及び評価データ入力部１２に入力された単一の新聞記事テキストに対してそれぞれ形態素解析処理を施し、これらの新聞記事テキストを単語へと分割し、第１の素性情報としての形態素解析結果を取得する。

【００６９】

文書長さ取得部１４は、学習コーパス保持部１１に保持されているすべての新聞記事テキスト及び評価データ入力部１２に入力された新聞記事テキストに対して、各テキストの長さ（テキスト中に含まれる文字数）を計測し、これを第２の素性情報として取得する。

【００７０】

学習コーパス選択部１５は、文書長さ取得部１４から得られるテキストの長さを基に、学習コーパスから、実際に機械学習で用いる学習データを選択する。ここで、Ｌは評価データ入力部１２に入力された新聞記事テキストの長さとし、Ｔはあらかじめ設定された正の定数とした場合、Ｌ－Ｔよりも長く、Ｌ＋Ｔよりも短いテキストを、学習コーパス保持部１１に保持されている新聞記事テキストから選択する。

【００７１】

学習コーパス選択部１５によって選択された新聞記事テキストを用いて、機械学習、並びに評価データ入力部１２から入力された新聞記事テキストの評価を行なう。本実施形態では、機械学習を文書分類システムに応用するが、Ｓｕｐｐｏｒｔ Ｖｅｃｔｏｒ Ｍａｃｈｉｎｅに基づいて、テキスト・データの素性と評価の間の対応規則を計算する。機械学習手段は、単語素性生成部２６と、文書素性生成部２７と、評価データ分類部２８で構成される。

【００７２】

単語素性生成部２６は、形態素解析部１３から得られるすべての単語に対して、対応する素性情報（集合）を生成する。以下、素性情報を生成するアルゴリズムについて説明する。

【００７３】

ステップ１：

形態素解析部１３から得られた全単語に対する表を作成する。但し、通常、新聞記事のキーワードとなりにくく且つ文に含まれる単語数の多い「は」又は「が」などの助詞については、ストップワードとして、表に入力しないことにする。

【００７４】

ステップ２：

形態素解析部１３から得られた単語をカウントし、ステップ１で得られた表に対して単語の個数を入力する。

【００７５】

文書素性生成部２７は、単語素性生成部１６で得られた素性情報を用いて、学習コーパス保持部１１中に保持されているすべての新聞記事に対応する素性情報を生成する。学習コーパス保持部１１は、複数の新聞記事を、記事毎に「政治経済」分野に属するか「スポーツ」分野に属するかを人手で判断した評価結果が入力されているが、データ形式は、上述した素性情報を生成するアルゴリズムと同等の方法で作成されている。ステップ２で得られた表を基に、学習コーパス保持部１１が保持する評価結果と比較して、文書素性情報を生成する。

【００７６】

例えば、学習コーパス保持部１１が保持する「政治経済」分野と「スポーツ」分野からすべての単語を抽出し、「政治経済」分野と「スポーツ」分野毎に得られた単語の表をそれぞれ作成する。「政治経済」分野と「スポーツ」分野から抽出された単語と一致するステップ２で得られた表に入力された単語のみ、これに対応する単語数を各分野の表に入力していく。これによって、「政治経済」分野と「スポーツ」分野の文書素性情報が生成される。カウントされなかった単語は、削除せず単語数０とする。

【００７７】

このようにして得られた文書素性は、学習コーパス保持部１１中の新聞記事集合を学習データとし、新聞記事に含まれる特徴表現語（ここでは「政治経済」分野と「スポーツ」分野の単語）を各記事の素性とした場合に得られる素性情報であると言える。

【００７８】

評価データ分類部２８は、「政治経済」分野と「スポーツ」分野の文書素性情報をＳｕｐｐｏｒｔ　Ｖｅｃｔｏｒ　Ｍａｃｈｉｎｅを用いた機械学習法を用いて計算させ、「政治経済」分野と「スポーツ」分野のいずれかに分類することが可能である。

【００７９】

図５には、本発明の第３の実施形態に係る機械学習システムの機能構成を模式的に示している。図示の機械学習システムは、学習データ保持部１と、評価データ入力部２と、第１の素性取得部３と、第２の素性取得部と、学習データ選択部５と、機械学習部６－１及び機械学習部６０２で構成される。この実施形態では、機械学習を行なう際に用いる学習データを選択するための第２の素性情報に対しＳｕｐｐｏｒｔ　Ｖｅｃｔｏｒ　Ｍａｃｈｉｎｅのような機械学習手法を適用し、得られた機械学習結果の中から精度の高い機械学習結果を用いて、学習データ選択のための素性取得に対応するルールを作成する。

【００８０】

第１の素性取得部３は、学習データ保持部１に保持されるデータ及び評価データ入力部２で受け取ったデータから、機械学習並びに評価を行なう際に用いる第１の素性情報を抽出する。

【００８１】

また、第２の素性取得部４は、学習データ保持部１に保持されるデータ及び評価データ入力部２で受け取ったデータから、機械学習を行なう際に用いる学習データを選択するた

めの第２の素性情報を抽出する。

【００８２】

機械学習部６－１は、第２の素性取得部４に対して、Ｓｕｐｐｏｒｔ Ｖｅｃｔｏｒ Ｍａｃｈｉｎｅのような機械学習手法を適用し、得られた機械学習結果の中から精度の高い機械学習結果を用いて、学習データ選択のための素性取得に対応するルールを作成する。そして、学習データ選択部５は、学習データ選択のための素性取得に対応するルールに従って、機械学習を行なう際に用いる学習データを選択する。

【００８３】

そして、機械学習部６－２は、学習データ選択部５によって選択された各学習データの評価と、第１の素性取得部３から得られた各データの素性を基に、素性とその評価の間の対応関係を計算する。

【００８４】

この実施形態では、学習コーパスから実際に機械学習を行なう新聞記事テキストを選択するための素性（上述した各実施形態では「テキストの長さ」）を用いたが、さらに品詞の並びなど文の特徴を追加して、複数の素性から適切な学習データを選択することも可能である。

【産業上の利用可能性】

【００８５】

以上、特定の実施形態を参照しながら、本発明について詳解してきた。しかしながら、本発明の要旨を逸脱しない範囲で当業者が該実施形態の修正や代用を成し得ることは自明である。

【００８６】

本明細書では、本発明に係る機械学習手法を新聞記事の分類（「政治経済」分野の記事であるか「スポーツ」分野の記事であるかの）などの文書分類システムに応用する場合を例にとって本発明について説明しているが、本発明の要旨はこれに限定されるものではない。すなわち、統計処理に基づく教師あり機械学習手法を用いるものであれば、アンケート分類及び質問応答など分類を要するあらゆる分野への応用であっても、同様に本発明を適用することが可能である。その他、テキスト分類のみならず数値データを含む分類や画像の分類など、いかなる機械学習手法を用いるものであっても、同様に本発明の効果を得ることが可能である。

【００８７】

要するに、例示という形態で本発明を開示してきたのであり、本明細書の記載内容を限定的に解釈するべきではない。本発明の要旨を判断するためには、冒頭に記載した特許請求の範囲の欄を参酌すべきである。

【図面の簡単な説明】

【００８８】

【図１】図１は、本発明に係る機械学習システムの機能構成を模式的に示した図である。

【図２】図２は、本発明の一実施形態に係る機械学習システムの機能構成を模式的に示した図である。

【図３】図３は、Ｓｕｐｐｏｒｔ Ｖｅｃｔｏｒ Ｍａｃｈｉｎｅを用いた機械学習における素性情報を示した図である。

【図４】図４は、Ｓｕｐｐｏｒｔ Ｖｅｃｔｏｒ Ｍａｃｈｉｎｅを機械学習に適用した場合の機械学習システムの機能構成を模式的に示した図である。

【図５】図５は、本発明の第３の実施形態に係る機械学習システムの機能構成を模式的に示した図である。

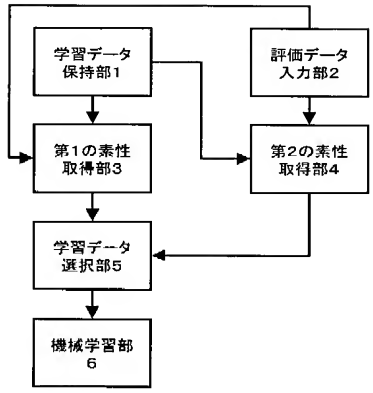
【符号の説明】

【００８９】

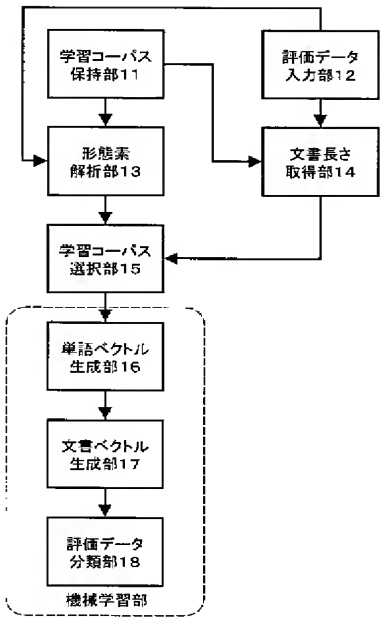
１…学習データ保持部
２…評価データ入力部
３…第１の素性取得部

4…第２の素性取得部
5…学習データ選択部
6…機械学習部
１１…学習コーパス保持部
１２…評価データ入力部
１３…形態素解析部
１４…文書長さ取得部
１５…学習コーパス選択部
１６…単語ベクトル生成部
１７…文書ベクトル生成部
１８…評価データ分類部
２６…単語素性生成部
２７…文書素性生成部
２８…評価データ分類部

【図１】



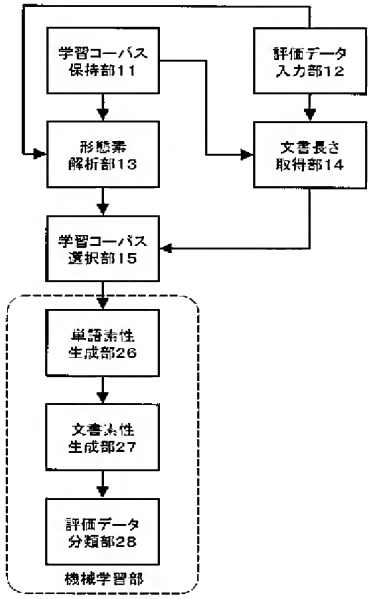
【図２】



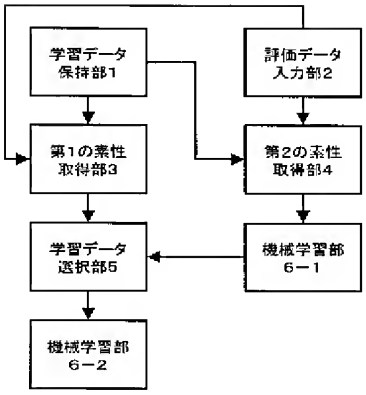
【図３】

	W_1	W_2	W_3	W_4	W_5	……	W_i
S_1	1	3	0	2	1	……	1

【図４】



【図５】



フロントページの続き

(72)発明者　大熊　智子
神奈川県足柄上郡中井町境４３０　グリーンテクなかい　富士ゼロックス株式会社内
(72)発明者　杉原　大悟
神奈川県足柄上郡中井町境４３０　グリーンテクなかい　富士ゼロックス株式会社内

Ｆターム(参考) 5B075 ND03　NK32　NR12

DERWENT-ACC-NO: 2005-452790

DERWENT-WEEK: 200546

TITLE: Machine-learning system acquires history information for performing learning and evaluation of received learning data, and acquires other history information to select learning data used when performing machine learning of received data

INVENTOR: MASUICHI H; OKUMA T ; SUGIHARA D ; YOSHIMURA H

PATENT-ASSIGNEE: FUJI XEROX CO LTD[XERF]

PRIORITY-DATA: 2003JP-426330 (December 24, 2003)

PATENT-FAMILY:

PUB-NO	PUB-DATE	LANGUAGE
JP 2005182696 A	July 7, 2005	JA

APPLICATION-DATA:

PUB-NO	APPL-DESCRIPTOR	APPL-NO	APPL-DATE
JP2005182696A	N/A	2003JP-426330	December 24, 2003

INT-CL-CURRENT:

TYPE	IPC DATE
CIPP	G06F17/30 20060101
CIPS	G06N3/00 20060101

ABSTRACTED-PUB-NO: JP 2005182696 A

BASIC-ABSTRACT:

NOVELTY - An acquisition unit (3) acquires history information for performing learning and evaluation of received learning data. An acquisition unit (4) acquires another history information for selecting learning data used when performing machine learning of received data. The learning data selected among several currently stored data based on other history information, is evaluated using the history information.

DESCRIPTION - INDEPENDENT CLAIMS are also included for the following:

(1) machine-learning method; and

(2) machine-learning program.

USE - For machine learning of document data related to sport field and politics-and-economics field.

ADVANTAGE - High precision machine learning is performed, using the appropriate learning data.

DESCRIPTION OF DRAWING(S) - The figure shows the block diagram of the machine-learning system. (Drawing includes non-English language text).

data retainer (1)

evaluation data input unit (2)

acquisition units (3,4)

selection unit (5)

CHOSEN-DRAWING: Dwg.1/5

TITLE-TERMS: MACHINE LEARNING SYSTEM ACQUIRE HISTORY INFORMATION PERFORMANCE EVALUATE RECEIVE DATA SELECT

DERWENT-CLASS: T01

EPI-CODES: T01-E05D; T01-J05B; T01-S03;

SECONDARY-ACC-NO:

Non-CPI Secondary Accession Numbers: 2005-368766